Object detection từ R-CNN đến Faster R-CNN

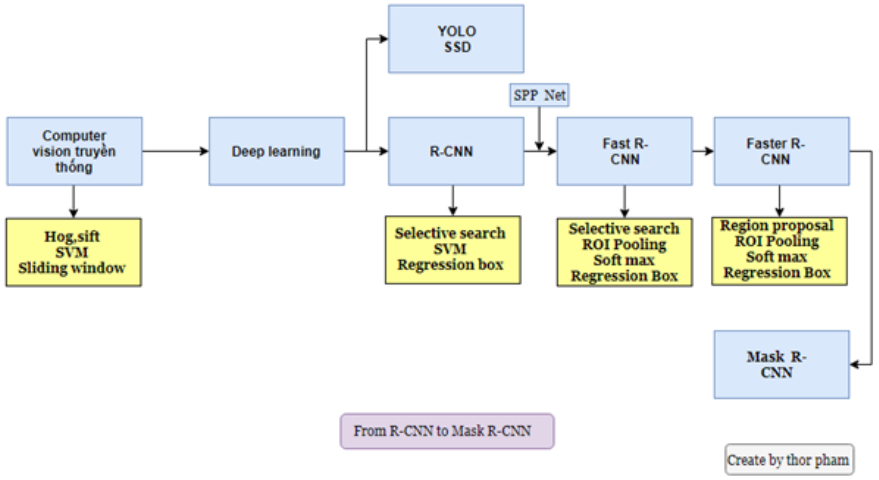
DEEP LEARNING VÀ ỨNG DỤNG · SATURDAY, JUNE 16, 2018

Như chúng ta đã biết object detection bao gồm 2 nhiệm vụ chính là **Classifier** và **Localization**. Trong đó nhiệm vụ có vẻ khó khăn hơn là Localization. Trước khi deep learning phát triển như hiện nay, trong computer vision người ta detection object qua 2 giai đoạn. Đầu tiên là trích xuất feature từ hog,lbp,sift sau đó dùng các thuật toán trong machine learning như SVM để classifier. Bước tiếp theo là detection object trên ảnh lớn thì người ta sẽ dùng 1 window search trên toàn bộ bức ảnh sau đó dùng model đã classifier để phân lớp object. Các model này có ưu điểm là thời gian build model tương đối nhanh, cần ít dữ liệu . Nhược điểm là độ chính xác không cao và thời gian predict rất lâu nên khó có thể dùng trong real time.

Với tộc độ phát triển như hiện nay , dữ liệu của chúng ta ngày càng nhiều và các bài toán bắt đầu khó dần lên nên những model truyền thống tỏ ra kém hiệu quả. Các feature lấy ra từ computer vision truyền thống như hog,sift,lbp là những shadow feature nó chỉ lấy được những feature trên bề mặt nổi image mà thôi, do đó những bài toán như classifer con chó hay con mèo thì những feature này làm việc tương đối hiệu quả , nhưng nâng cấp bài toán lên đó là classifier con bull dog hay con béc rê thì những feature này làm việc kém hiệu quả. Cũng dễ hiểu vì cùng 1 class là dog thì những feature này tương đối giống nhau nên khó có thể phân biệt được con này con kia. Chính vì thế ta cần những feature sâu hơn, những feature mà nó ẩn ở trong image mà ta khó có thể quan sát được bằng mắt thường để phân biệt dog này hay dog kia. Trong deep learning , object detection nền tảng cơ bản là dựa trên mạng CNN để lấy những deep feature bằng cách đưa qua nhiều layer khác nhau feature được extract sâu hơn sau đó được đưa vào classifier và regression box. Hai hướng tiếp cận chính trong deep learning là :

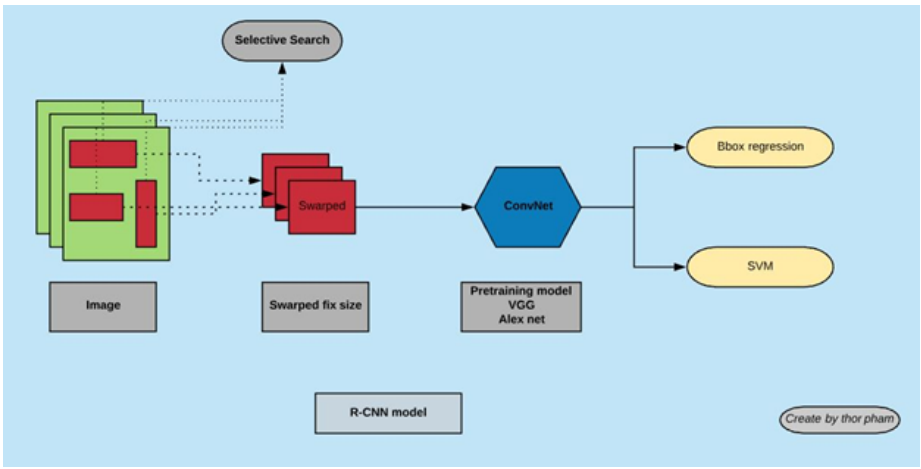
- Chia image ra thành những grid cell SxS . Mỗi cell được coi như region proposal giúp giảm thời gian và chi phí tính toán thay vì sử dụng trực tiếp image (model SSD,YOLO)
- Tìm những region proposal có nhiều khẳn năng chứa object nhất sử dụng selective search hay RPN (model R-CNN,Fast R-CNN, Faster R-CNN)

Trong bài này chúng ta sẽ tìm hiểu về các họ nhà CNN cho object detection.



Map of object detection

R-CNN



R-CNN

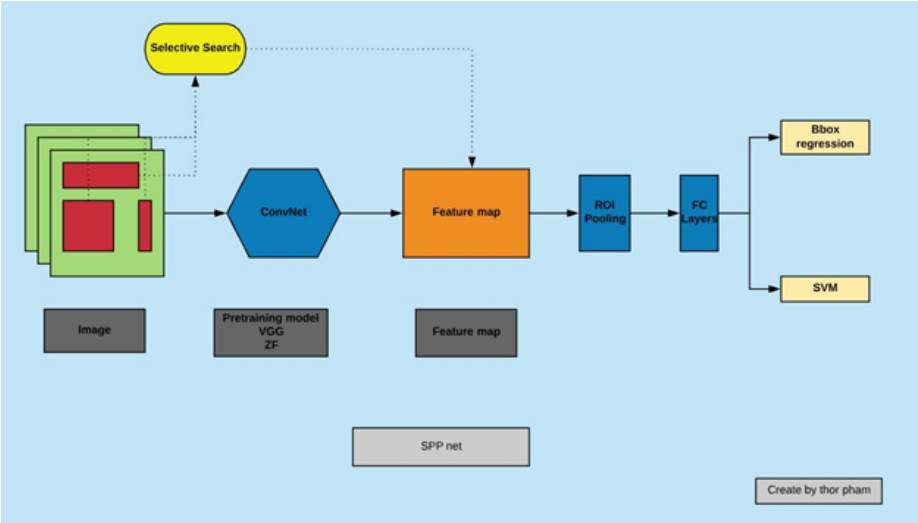
R-CNN là viết tắt của “Region-based Convolutional Neural Networks”. Ý tưởng chính của nó gồm 2 phần : Đầu tiên là dùng selective search tìm các region of interest (roi) trên image để tạo ra các bounding box mà có xác suất cao là có object. Sau đó dùng CNN để lấy feature từ các region này để classifier và regression box. Các bước thực hiện như sau :

1. Từ input image ta dùng selective search để lấy cái region proposal .Selective Search nó hoạt động bằng cách là đầu tiên tạo ra các seed là các region segementation trên image(trong skimage họ dùng Felsenszwalb’s efficient graph based image segmentation) Các region sau đó sẽ được merger lại với nhau bằng cách tính độ tương đồng về color,shape,textuture… Cuối cùng ta vẽ bounding box cho từng region.Mỗi image người ta sẽ lấy tầm 2k region proposal.
2. Tiếp đến các region proposal sẽ được swarped(crop) để fix size(vì một số mạng như VGG yêu cầu size input là cố định,hơn nữa ta cần feature output same size). Sau đó dùng CNN(VGG,Alexnet) để lấy feature.

3. Cuối cùng là dùng SVM để classifier và regression bounding box

Nhược điểm của phương pháp này là training rất lâu vì 2k image qua CNN để lấy feature mất rất nhiều thời gian (52 s trên cpu) .

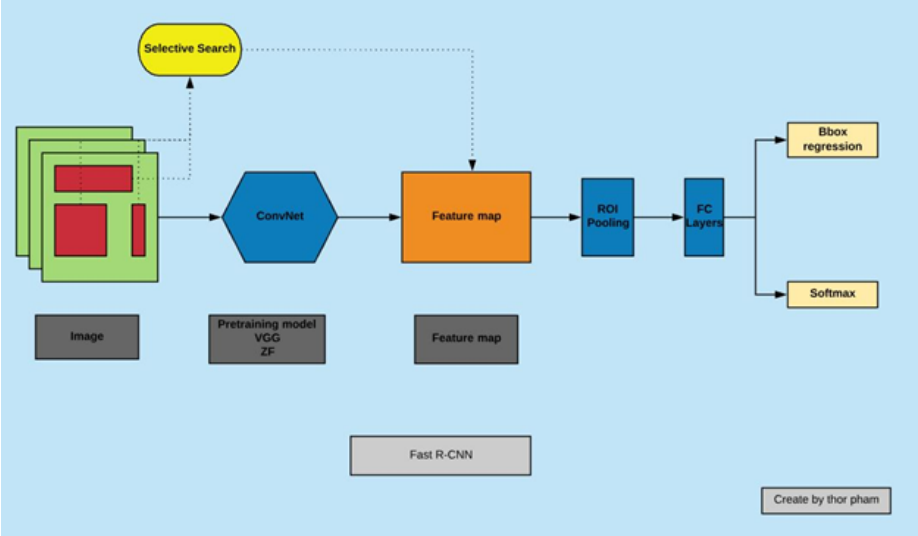
SPP net



SPP net

Thay vì feed forword 2k image qua CNN thì người ta feed forwork qua CNN một lần để lấy feature map (feature map = feature + location). Sau đó dùng selective search để tìm region proposal, rồi project trên feature map để lấy feature tương ứng. Có một vấn đề ở đây là các feature map của region proposal có size khác nhau nên khi đưa qua CNN sẽ có length output khác nhau. Vì vậy người ta sử dụng Spatial paramy pooling layer để fix size feature.(spp layer hoạt động cũng tương tự bag of word trong image processing nó sẽ chia feature theo Spatial pyramid và áp dụng max pooling theo từng spatial giúp các feature có size khác nhau thành same size). Phần sau còn lại tương tự như R-CNN

Fast R-CNN



Fast R-CNN

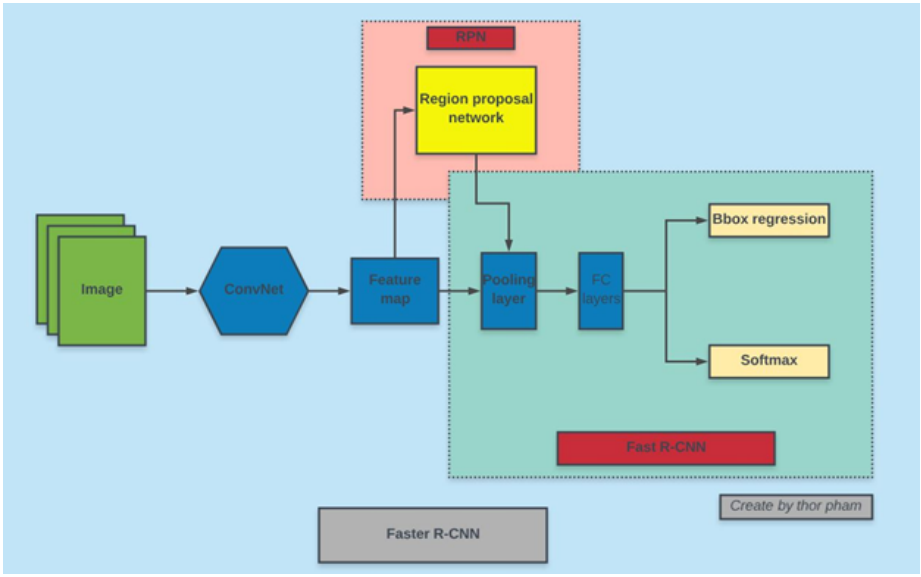
Fast R-CNN cải thiện được các nhược điểm của R-CNN bằng cách hợp nhất 3 model độc lập vốn rất chậm chạp. Nó cũng có phần giống SPP-net là dùng CNN để lấy feature map một lần thay vì dùng riêng cho mỗi region proposal. Sau đó những feature này sẽ được đưa qua một Fully connection layer để classifier và regression bounding box. Model này tương đối nhanh chạy 2s/image.

Các bước thức hiện thuật toán :

Dùng pretraining model (VGG,ZF...) để lấy feature map.

1. Sử dụng selective search để lấy region proposal (~2k image). Sau đó project lên feature map để lấy feature tương ứng
2. Feature sẽ được đưa qua ROI pooling để fix size.
3. Cuối cùng Fully connection layer để classifier và regression box

Faster R-CNN



Faster RCNN

Faster R-CNN gồm region proposal network (nó thay thế cho selective search) và phần còn lại tương tự như Fast R-CNN. Region proposal network(RPN) nó dùng 1 sliding window search trên feature map để tạo các anchor box. Sau đó chúng ta chuẩn bị data training cho RPN bằng cách gán nhãn cho mỗi anchor box dựa vào iou với ground truth. Cuối cùng dùng data này để classifier và regression bounding box. Ta thu được rất nhiều bounding box và dùng non maximum suppression(NMS) để loại bỏ bớt đi những box không có nhiều khẳn năng chứa object. Sau đó những bounding box này sẽ tương tự như selective search ở fast R-CNN nó được đưa qua ROI pooling để fix size và cuối cùng đưa vào fully connection layer để classifier (xác định object cụ thể) và regression box. Model này tương đối nhanh predict 0.2s/image(gpu)

Model gồm các bước sau :

Pretrain model CNN để lấy feature map.

1. Training RPN để tìm bouding box và classifier (chỉ xác định là object và non-object không classifier cụ thể object). Một siliding window size NxM search trên feature map. Tại mỗi center của window, ta predict mutil region với scale và ratio khác nhau. Thông thường là 3 scale và 3 ratio nên tạo ra 9 anchor box. Positive sample IOU > 0.7, negative sample IOU < 0.3.
2. Dùng data này training để classifier và regression . Vì số background nhiều nên để hạn chế bias người ta dùng mini bath để training mỗi lần đưa vào tỉ lệ một pos và neg nhất định. Sau đó loại bớt bounding box có ít khẳn năng chứa object bằng NMS.
3. Ta project bounding box lên feature map để lấy feature tương ứng sau đó đưa vào ROI pooling layer fix size để đưa vào Fully connection layer để classifier từng object và regression box.

Kết luận : Trên đây mới chỉ là một bài giới thiệu sơ qua về cách hoạt động của R-CNN đến Faster R-CNN. Bên trong cấu trúc của mỗi model này tương đối phức tạp. Nếu có thể mình sẽ viết chi tiết cách hoạt động của từng model qua các bài sau.